蕎麥葉片及種子轉錄體分析

陳裕星、王捷、陳鐶斌、廖宜倫
行政院農業委員會台中區農業改良場

摘　要

本研究分析蕎麥種子及葉片轉錄體，分別讀取 143 M 及 247 M 條序列，共組裝 97,200 及 74,544 條 contigs，平均覆蓋率(coverage)分別約為 1,474 及 3,340 倍，contig 序列平均長度分別為 565 及 607 bp，全部拼接完成之序列分別為 54M 及 45M bp。葉片及種子序列混合拼接時，共可獲得 106,190 條 contig，其中屬於種子和葉片者分別為 69,043 及 56,345 條，單獨存於種子或葉片者分別為 22,587 及 9,889 條。所有組裝成功之 contigs 進一步與資料庫比對，註解並進行基因功能分類檢索，分別依照細胞組成 (cellular component)、生理程序 (biological process) 及分子功能進行分類。本研究同時檢索與種子蛋白質生合成相關基因，包括種子貯存蛋白、LEA protein, cupin 家族蛋白及過敏蛋白等基因，這些資訊可提供未來蕎麥栽培育種工作之參考。

關鍵詞：芸香苷、過敏蛋白、蕎麥核心抗原、蕎麥糖醇。

前　言

蕎麥（*Fagopyrum esculentum* Moench）屬於蓼科蕎麥屬一年草生作物，蕎麥種子與植株中含有多種有益健康之機能性成份，同時為優良蜜源植物，因此成為世界重要雜糧作物之一。屬內有許多種，最常見的栽培種有二種，一為普通種蕎麥 (common buckwheat) *Fagopyrum esculentum* Moench，另一為韃靼種蕎麥 (Tatary buckwheat) *Fagopyrum tataricum* (L.) Gaertn。蕎麥原為台灣中部地區水田秋冬裡作作物，主要供做綠肥用(曾及陳，2007、2008)，在台灣主要之栽培地區以彰化二林、大城等地為主，面積約 40 公頃，於二期作水稻收穫後秋裏作種植，台中區農業改良場目前已育成「台中 1 號」、「台中 2 號」、「台中 3 號」及「台中 5 號」等品種，並推廣給農民栽培，其中台中 2 號為韃靼種蕎麥，台中 5 號為普通種蕎麥(曾等，2009)，為目前主要栽培品種。普通種蕎麥同時為優良之蜜源植物，盛花期可長達 2-3 個月，每公頃並可產蜜達 70-100 公斤 (Cawoy et al., 2008)，在農村可搭配蕎麥粉手工製麵，作為副產業與農村休閒旅遊之用。

本研究針對蕎麥種子及葉片轉錄體，進行基因功能分類 (gene ontology) 與註解 (annotation)，對比葉片及種子基因表現的輪廓，並探討同一物種葉片與種子基因表現差異之情形，藉此了解對產量及品質相關重要基因的表現情形，包括澱粉合成酶、種子貯存蛋白質、過敏蛋白質的種類及表現情形。

材料與方法

一、植物材料

蕎麥台中 5 號於 2013 年 1 月以種子條播方式定植於台中區農業改良場(彰化縣大村鄉)，定植後第 6 週即陸續開始開花結實，於

第 7 週收穫充實期種皮為綠色之種子，以及與同一植株最上方之成熟葉片，採回實驗室後立即進行 RNA 之萃取。

二、RNA 萃取

種子與葉片 RNA 之萃取使用 TANBead® Plant RNA kit，首先取 50 mg 種子及葉片組織各 4 重複，單一樣品置於小試管中，加入試劑組之 lysis buffer 後以植物組織均質器研磨，之後取出均質液轉到 TANBead 全自動磁珠操作平台(台灣圓點奈米技術公司)，自動萃取純化 RNA，所純化之 RNA 暫存於 -20℃冰箱保存。

三、轉錄體定序

種子與葉片之 RNA 樣品使用 Agilent Bioanalyzer 2100 以 RNA 6000 進行品質分析，以 RNA 完整性(RNA Integrity Number, RIN)值 7.0 為閾值，棄卻 RIN 低於 7.0 之樣品，本研究所採用之種子與葉片樣品 RIN 值分別為 8.7 及 7.3。RNA 樣品使用 Nugen Ovation RNA Seq System V2 進行反轉錄反應合成 cDNA；cDNA 以 Covaris S220 剪切用以製備固定長度 cDNA 之文庫構築，使用 Nugen Encore Rapid DR system 將 cDNA 兩端接上接合器後，再經過純化、定量與文庫長度分布確認後，隨即以 Illumina HiSeq 2000 上機進行雙邊各 100 bp 定序(如圖 1 工作流程)。

四、轉錄體分析與基因分類註解

以 CLC Genomics workbench 套裝軟體之“trim 模組”去除訊號不良之序列，並使用”de novo 組裝模組”將種子與葉片之 reads 混合組裝 (由於無法獲得蕎麥詳細的基因組資訊故以此作為本次參考基因訊息)；以 blastx 與 NCBI nr 資料庫比對，以 FastAnnotator (Chen et al., 2012)進行基因註解(annotation) (圖 2)。再將此註解好之 contigs 當作參考基因組，分別將種子與葉片之 reads 進行 RNA-Seq 定序，並以 RPKM

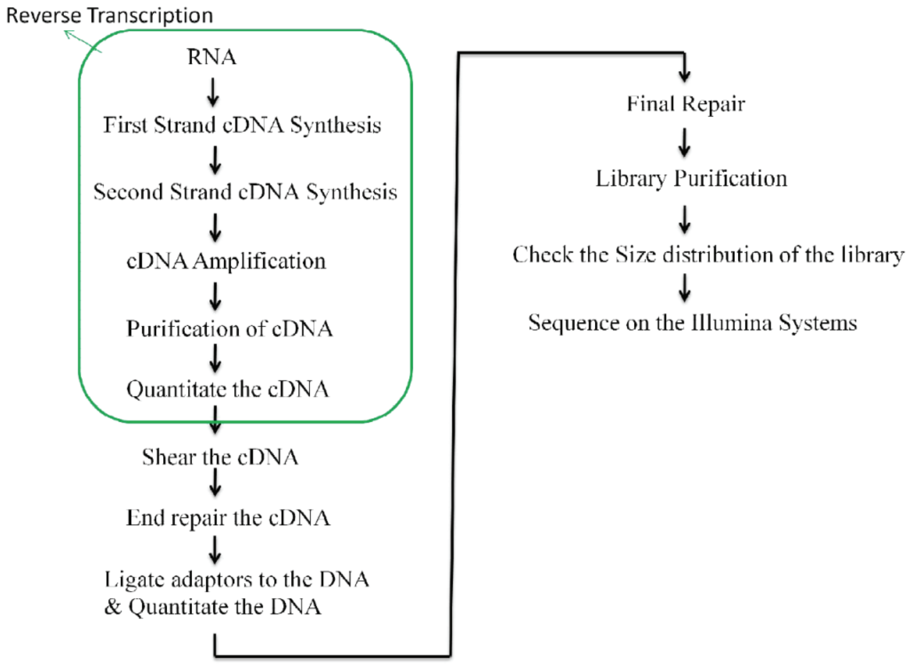

圖 1. 蕎麥種子與葉片轉錄體材料的製備流程。

Fig. 1. Preparation of seeds and leaves transcriptome of buckwheat.

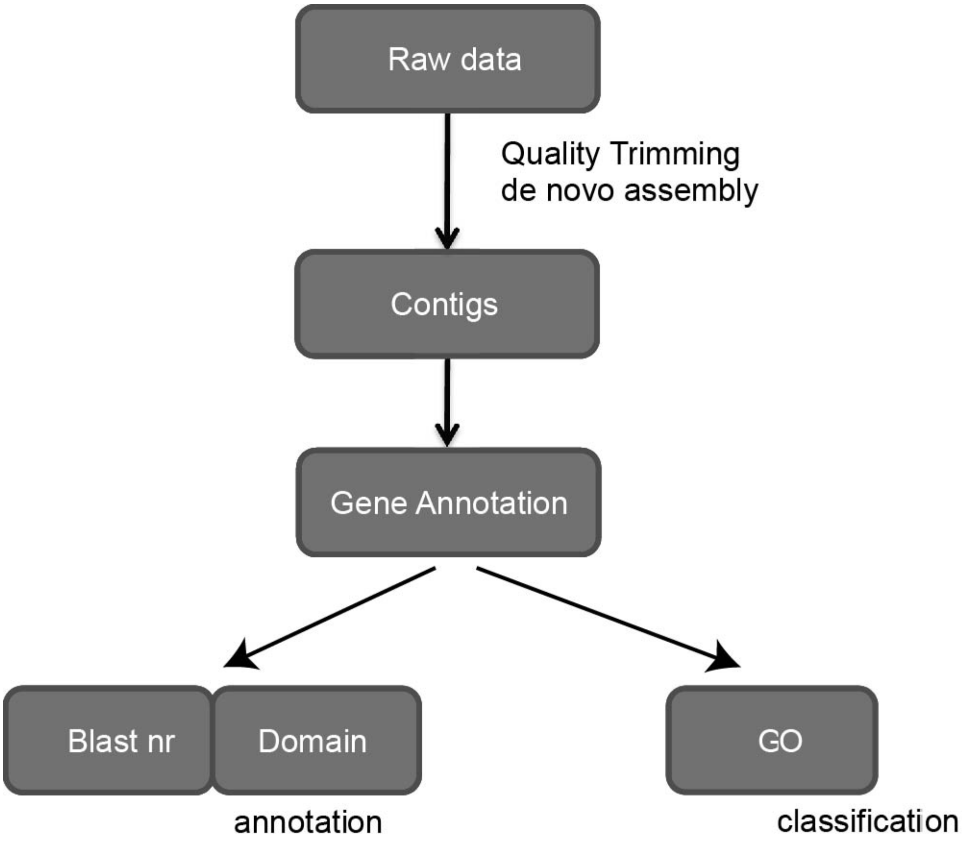

圖 2. 轉錄體分析與基因分類註解流程。
Fig.2 Flow chart of gene ontology classification and annotation.

(Reads Per Kilo-base per Million reads)來表示參考基因之表現量(Wagner et al., 2012)，將種子與葉片表達量有差異之 contigs 比較分析。

結果與討論

本研究中送定序之種子(BS)與葉片(BL)樣本分別讀取 143 M 及 247 M 條序列，共組裝 97,200 及 74,544 條 contigs，平均覆蓋率(coverage)分別約為 1474 及 3340 倍，contig 序列平均長度分別為 565 及 607 bp，全部拼接完成之序列分別為 54M 及 45M bp (表 1)。參考 Logacheva 氏等(2011) 進行之蕎麥花朵轉錄體分析，共讀取 266,782 條序列，獲得 25,435 條 contig，平均長度為 698 bp，覆蓋率為 8.2。

本研究所使用之種子及葉片轉錄體樣品同為蕎麥臺中 5 號，完成種子及葉片轉錄體分析之後，進一步混合拼接以比較種子和葉片表現基因之異同，如此可以得知哪些 contigs 分別屬於種子或葉片，或是屬於兩者共有，不屬於葉片或種子則可視為錯誤拼接(圖 3C) 可排除之。其次，混合拼接可將不完整的同源序列組裝為更完整之 contigs。

所有讀取之序列進行混合拼接，共可獲得 106,190 條 contig，其中屬於種子和葉片者分別為 69,043 及 56,345 條 (圖 3)，單獨存於種子或葉片者分別為 22,587 及 9,889 條；混合拼接後衍生 27,258 條(圖 3C 區) contig 其 RPKM 值為 0，這些組裝後無法算出 RPKM 值的 contig，表示其為由兩者共同 assembly

表 1. 原始資料及重疊群 (contig) 之特徵
Table 1. Characteristics of raw data and contigs

	Buckwheat seeds	Buckwheat leaf	BS-BL mixed assembly
Number of reads	143,803,530	247,195,816	390,999,346
N75	370	397	445
N50	696	810	840
N25	1,457	1,651	1,716
Minimum Length	135	68	106
Maximum Length	17,000	14,726	23,106
Average Length	565	607	640
Number of Contigs	97,200	74,544	106,190
Total bps	54,889,512	45,283,362	67,945,620

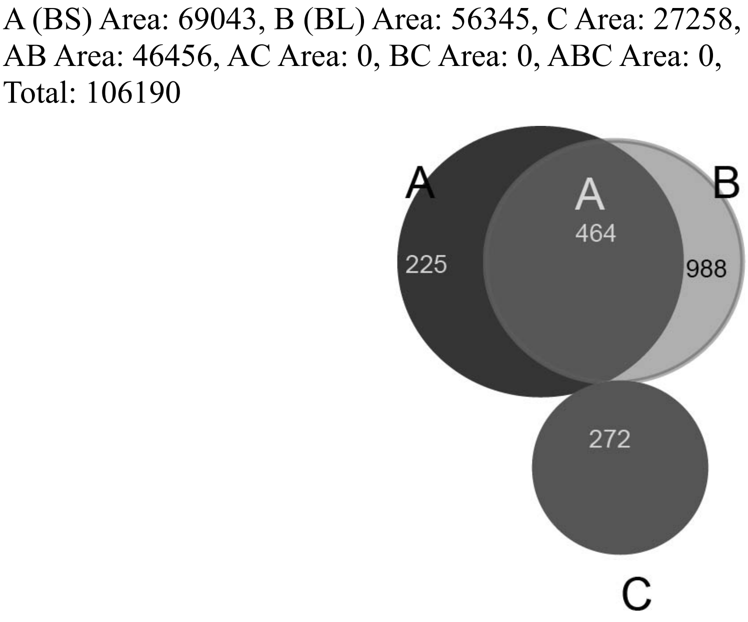

圖 3. 蕎麥葉片與種子個別拼接與混合拼接後所獲得 contig 之片段分類。
Fig. 3. Contigs segmentation after mixed assembly.

出的 contig，不屬於種子或葉片，在此認定為錯誤之拼接，佔總 contigs 數量的 25.6%，由原始資料也可以看出這些共同組裝的 contig 其長度平均也短，為 260 ± 53 bp，進行 blast nr 比對結果，大多也無法得到理想的結果 (資料未顯示)，推測為降解的 RNA。扣除錯誤拼接的 contigs 之後，成功組裝的 contigs 認定為 78,932 條 (106,190 – 27258 = 78932)。

以混合拼接後各個 contig 在種子和葉片中標準化後的 RPKM 值繪製散佈圖。比較相同 contig 在兩種子及葉片中相對表現差異 (圖 4)，整體而言，同樣的 contig 在種子中有較高的表現量。

進一步分析各 contig 之 RPKM 值可發現，種子和葉片中各有 37,147 及 49,845 條 contigs 之 RPKM 值趨近於 0，所比對到序列長度(hit length)也都極低 (15~100 bp)，RPKM 值介於 0-1 之間者各有 33,178 及 35,287 條 (圖 5)，這些 contigs 因為表現量低，或是因為錯誤組裝使得其在整體資料中相對為微量，或者是因為其本身表現即為微量，推測並非我們有興趣之基因。蕎麥引人注目的機能性成分包括芸香苷、手性肌醇 (D-chiral inositol)、cupin 家族的過敏蛋白質等，從文獻得知這些成分含量在蕎麥中都相當高，本研究中專注在高表現量的 contig 分析；在數據的處理上，篩選掉種子 RPKM<10 之後的 contig，總共剩下 8,973 條 contig (圖 5)。此外，混合拼接之後表現量高於 1 的 contig 在種子樣品中數量也較多 (圖 5)，因此本研究篩選在種子中具有高表現量 (RPKM>10)，且和葉片表現量相比也有高比值者優先進行分析。

所有 contig 序列與資料庫相比對，其中 33,342 筆可比對出基因功能 (gene ontology)，與蛋白質區塊資料庫比對有 21,649 筆，與酵素資料庫比對則可得到 2,580 筆 (圖 6)，而其中三者交集的序列為 2,157 筆，三者聯集者為 37,936 筆。整體而言所有比對到的筆數占全部認定為成功組裝 contigs 的 48.1%。

依照細胞組成加以分類，佔比重最高的前五名依序為細胞膜成分 (integral to membrane)、原生質膜 (plasma membrane)、細胞質 (cytosol)、細胞核 (nucleus)以及粒線體 (mitochondria) (圖 7)。前二名皆屬於細胞膜上之功能基因，可見膜功能在生物中的重要性，受限於時間因素，目前尚未和其他植物文獻報告比對，而 Logacheva 氏等 (2011) 所發表的蕎麥花芽轉錄體並未提出以細胞組成分類的結果。

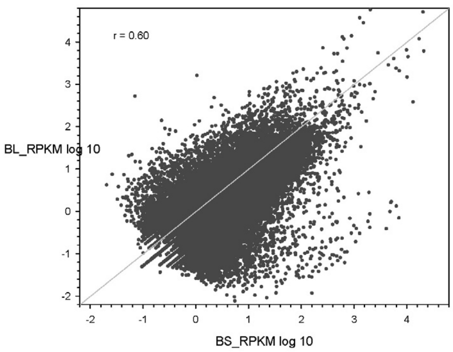

圖 4. 種子和葉片標準化後 contig 之 RPKM 數值散佈圖

Fig 4. Scattered plot of standardized seed and leaf RPKM value.

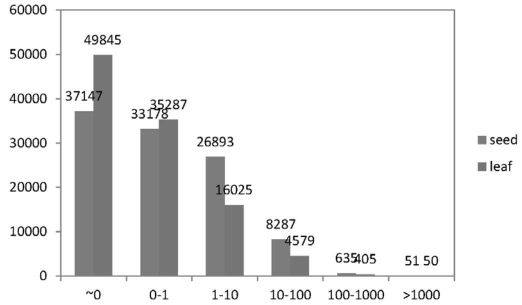

圖 5. 蕎麥葉片與種子混合拼接之 contig 依照 RPKM 值分類圖

Fig. 5. Distribution of RPKM value of mixed assembled contigs.

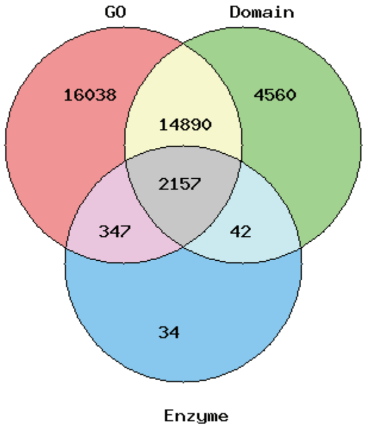

圖 6. 蕎麥種子與葉片轉錄體註解分類

Fig. 6. Buckwheat contigs (BS-BL) annotation segmentation

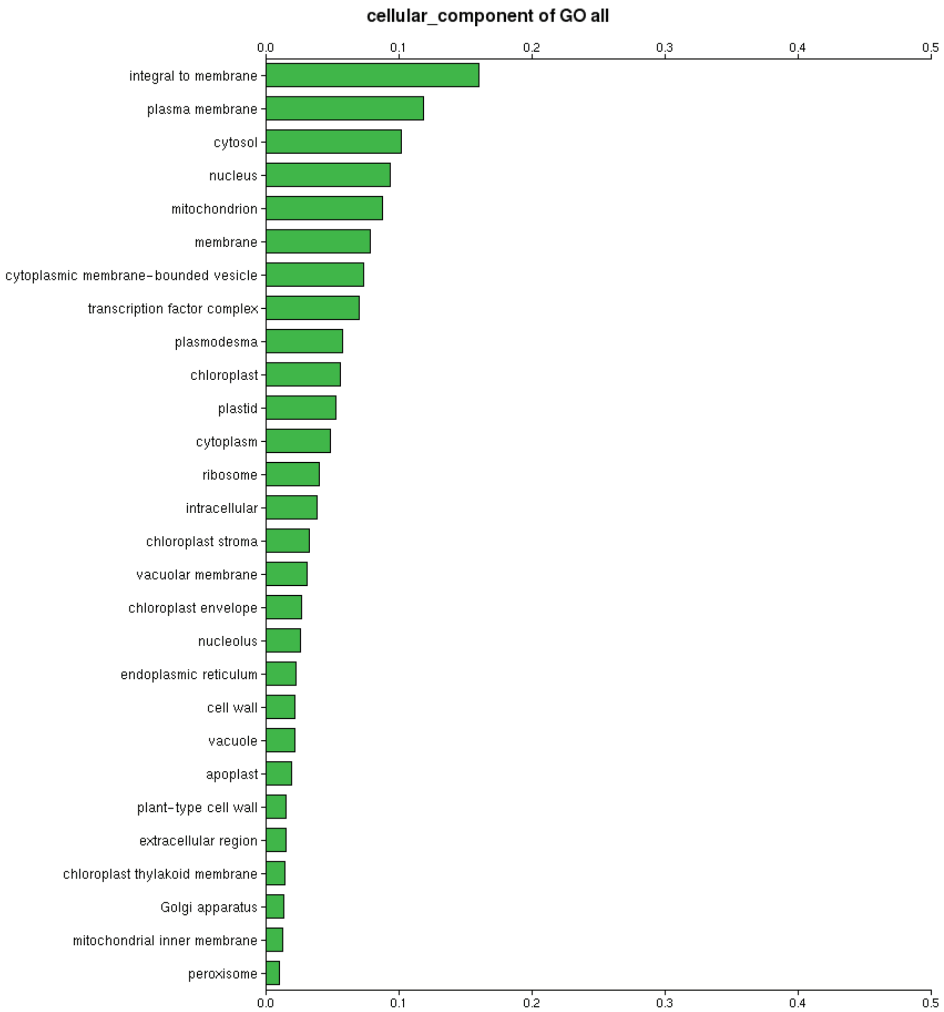

圖 7. 以細胞成份為標準之基因功能分類。

Fig. 7. Gene ontology classification according to cellular components.

依據 Logacheva 氏等分析蕎麥花芽轉錄體的報告，其使用簡化的 GOSlim 註解方式和阿拉伯芥的生物資訊 (TAIR) (Rhee et al., 2003) 互相比較，依照生理程序分類蕎麥花芽轉錄體之五項分別為生理程序、細胞程序、代謝程序、初級代謝程序與細胞程序，可知此分類方式相當簡化。本研究比對 2013 年之 NCBI 資料庫序列並註解，生理程序排序前五名分別為氧化還原反應、絲胺酸家族胺基酸代謝、DNA 依賴性的轉錄控制、蛋白質磷酸化以及電子傳遞鍊 (圖 8)，

當以分子功能為分類依據時，排序前五名分別為 ATP binding, Zinc ion binding, DNA binding, sequence-specific DNA binding transcription factor activity, and binding (圖 9)。而 Logacheva 氏等 (2011) 之報告其排序前五名分別為 molecular function, binding, catalytic activity, transferase activity and protein binding，與本報告分類方式不同，但是並無太大之牴觸。

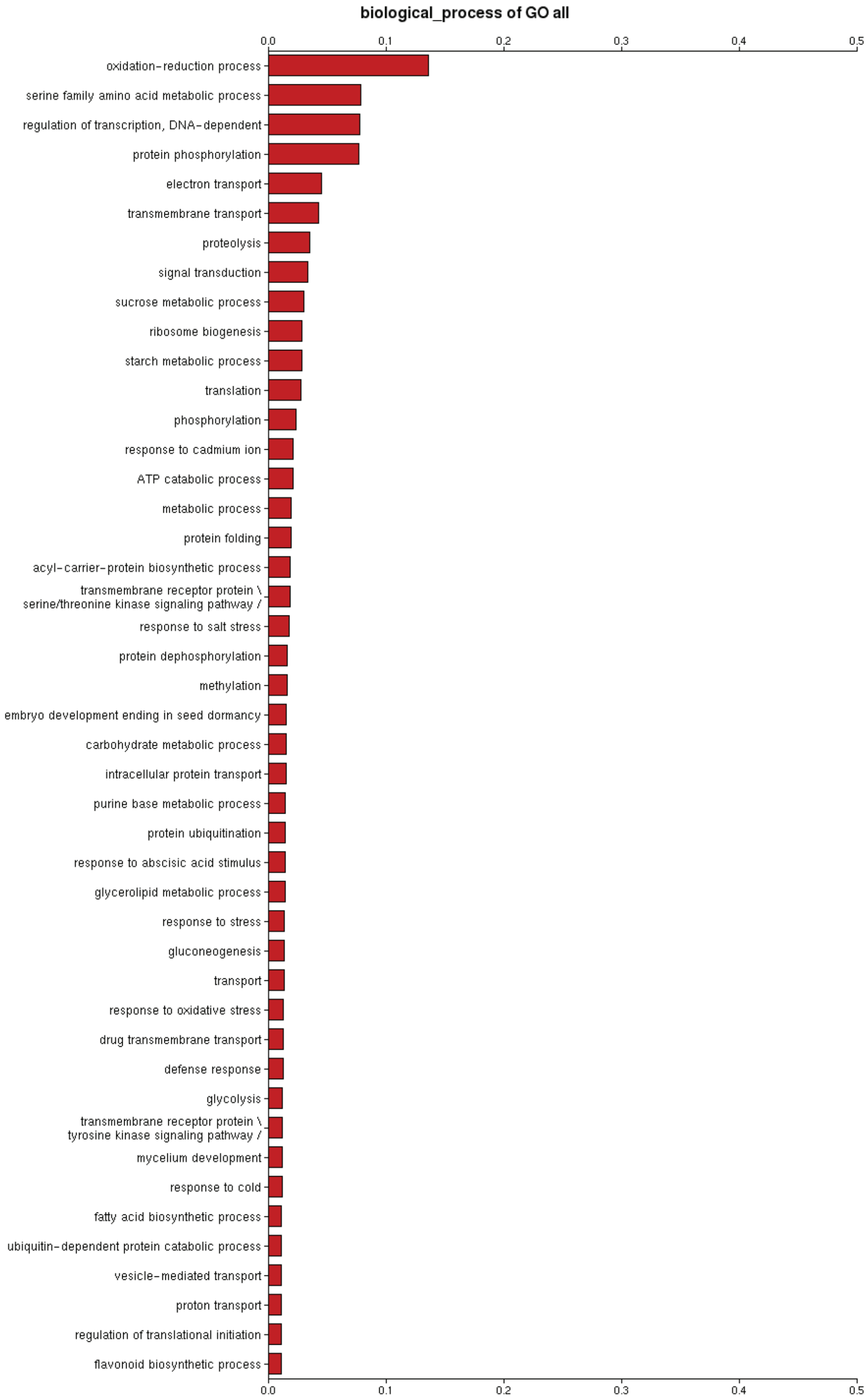

圖 8. 以生理程序為標準之基因功能分類。

Fig. 8. Gene ontology classification according to biological processes.

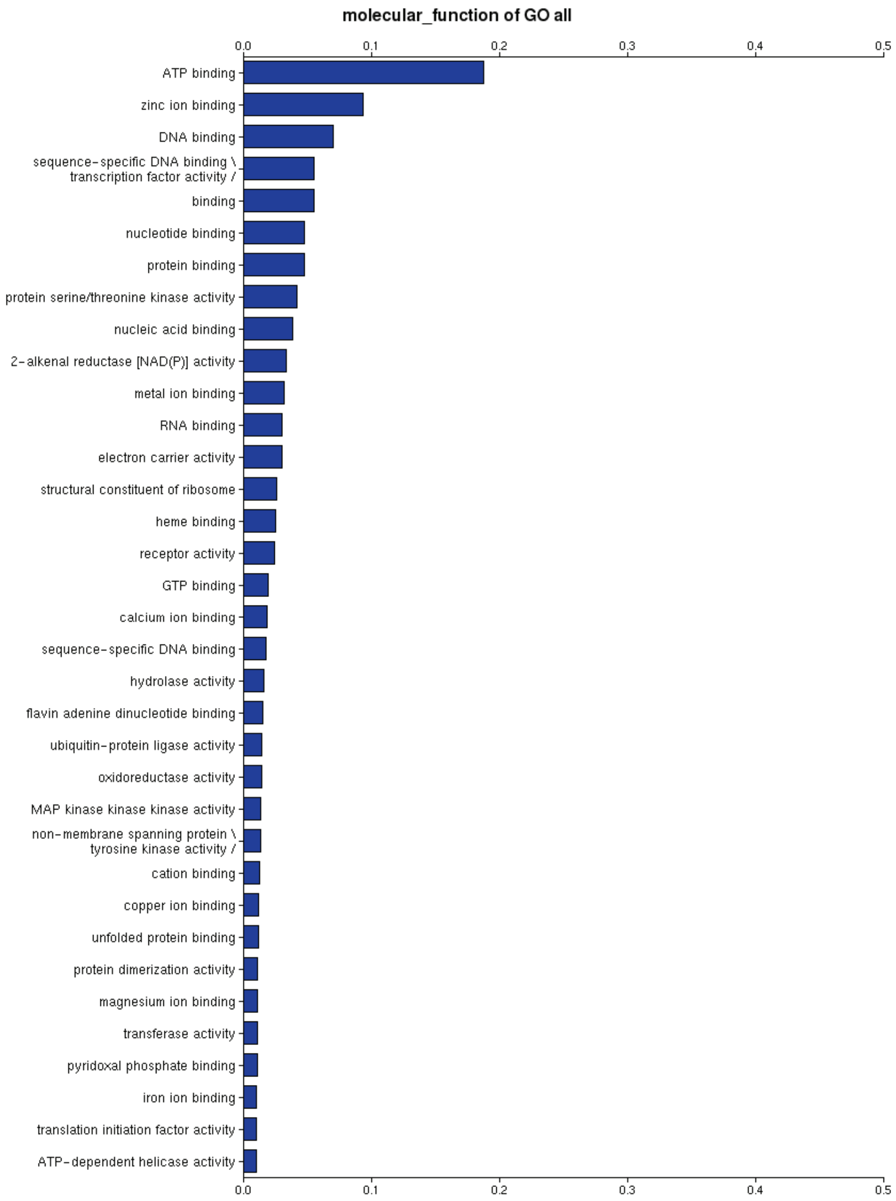

圖 9. 以分子功能為標準之基因功能分類。

Fig. 9. Gene ontology classification according to molecular functions.

種子中高表現量 contigs 之篩選

和農業栽培最相關者為種子產量與品質，我們認為特定 contigs 在種子中表現量高 (RPKM>10)，即為與種子充實高度相關之基因。同時，比較在種子和葉片表現量比值，比值高者為該基因在種子的表現量高於葉片，如此即可得知，有多少數量的 contig 在種子中有專一性的大量表現 (圖 6)。由圖 10 之 BS/BL RPKM 比值可知，共有 7,822 條 contig 在種子有高表現量，未來將優先探討這些高表現量 contigs 的基因功能。

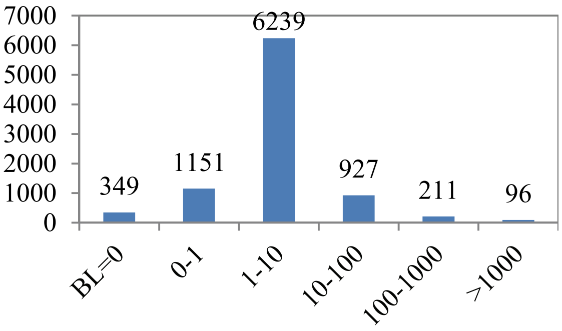

圖 10. 蕎麥種子和葉片 BS/BL contigs RPKM 比值之分佈

Fig. 10. Distribution of the BS/BL contigs RPKM ratio

結論

由於次世代定序技術(Next Generation Sequencing, NGS)的快速發展，使得基因體與轉錄體的分析得以快速進行，近年來學界已經累積豐富的作物生物資訊，科研瓶頸反而出現在資料的解析。本研究雖已完成蕎麥葉片與種子轉錄體之定序與註解，然而分析工作仍在早期階段，目前僅能歸納本轉錄體大致的輪廓，尚未進入深入分析階段。蕎麥中有數項引人入勝的議題，包括植株可以合成累積超高量的芸香苷 (Kusano et al., 1996; Li et al., 2001; Fabjan et al., 2003)，可達 46,348 ppm，相當於 4.6 %的乾物重，筆者在分析本場所栽培之蕎麥台中 5 號時獲得此數據，與國外所發表文獻相當。蕎麥種子含有豐富的手性肌醇(obendorf, 1998; Steadman et al., 2000)，為胰島素訊息傳導受體 PIP2 的組成成分之一 (Baillargeon et al., 2010)，可增加糖尿病第 II 型病患對胰島素之敏感性，也有助醫治 PCOS 病患之病徵 (Nestler, et al., 1999, 2000)。DCI 與 galactopyranose 合成 fagopyritol A2、A3、B2、B3，為麩皮上之可溶性碳水化合物 (Steadman et al., 2000)，可開發為具醫療效果的膳食補充品。然而蕎麥亦具有過敏性蛋白質 (Wieslander, 1995)，雖然蕎麥所造成食物過敏現象的盛行率遠低於小麥麩質、海鮮、花生堅果等八大類常造成過敏食品 (Nair and Adachi, 1999)，若能透過育種方式降低蕎麥過敏蛋白質含量，對蕎麥產業會有極大的幫助。本研究分析蕎麥種子和葉片轉錄體，報告中針對學術界對於蕎麥有興趣的幾項議題進行初步篩選分析，未來仍需更深入研究影響蕎麥產量與品質的因素，以期對蕎麥栽培育種及加工有更大的助益。

誌　　謝

本研究由農委會科技計畫補助 (計畫名稱：蕎麥及薏苡 DNA 生物條碼與化學指紋圖譜分析模式之建立，計畫編號：102 農科-9.1.1-中-D1)，轉錄體資訊在中研院生物多樣性中心趙淑妙老師指導下進行，並得莊樹諄老師提供分析建議，與研究室其他學弟妹共同討論轉錄體分析訣竅，源資生物科技公司許宏吉博士協助資料剖析，特此致謝。

引用文獻

曾勝雄、陳裕星 2007 蕎麥台中 2 號之育成 台中區農業改良場研究彙報 95: 49-59。

曾勝雄、陳裕星 2008 蕎麥台中 3 號之育成 台中區農業改良場研究彙報 100: 13-23。

曾勝雄、陳裕星、廖宜倫 2009 蕎麥台中 5 號之育成. 台中區農業改良場研究彙報 103: 1-15。

Baillargeon J. P., M. J. Iuorno, T. Apridonidze and J. E. Nestler. 2010. Uncoupling between insulin and release of a D-chiro-inositol-containing inositolphosphoglycan mediator of insulin action in obese women With polycystic ovary syndrome. Metab Syndr Relat Disord. 8(2):127-36.

Cawoy, V., J. M. Kinet and A. L. Jacquemart. 2008. Morphology of nectarines and biology of nectar production in the distylous species *Fagopyrum esculentum*. Annals of Botany 102: 675-684.

Chen T. W., R. C. R. Gan, T. H. Wu, P. J. Huang, C.

Y. Lee, Y. Y. M. Chen, C. C. Chen and P. Tang. 2012. FastAnnotator: an efficient transcript annotation web tool. BMC Genomics 13:Suppl 7, S9.

Fabjan N., J. Rode, I. J. Kosir, Z. Wang, Z. Zhang and I. Kreft. 2003. Tartary buckwheat (*Fagopyrum tataricum* Gaertn.) as a source of dietary rutin and quercitrin. J Agric Food Chem. 51:6452-6455.

Kusano T., N. Nakayama, J. Doi, Y. Kiyomasa, H. Chine and K. Ikeda. 1992. Studies on nutritional aspects of tartary buckwheat. In Proc. 5th Intl. Symp. Buckwheat.; 453-454. August 20–26 1992, Taiyuan, China.

Li S. Q. and Q. H. Zhang. 2001. Advances in the development of functional foods from buckwheat. Crit Rev Food Sci Nutr. 41:451-464.

Logacheva D., A. S. Kasianov, D. V. Vinogradov, T. H. Samigullin, M. S. Gelfand, V. J. Makeev and A. A. Penin. 2011. *De novo* sequencing and characterization of floral transcriptome in two species of buckwheat (*Fagopyrum*). BMC Genomics 12: 30.

Nair A. and T. Adachi. 1999. Immunodetection and Characterization of Allergenic Proteins in Common Buckwheat (*Fagopyrum esculentum*). Plant Biotechnology, 16 (3), 219- 224.

Obendorf R. L. 1998. Buckwheat Fagopyritols. Advances in buckwheat research. 7th Intl Symp Buckwheat. 3:65-71.

Rhee S. Y., W. Beavis, T. Z. Berardini, G. Chen, D. Dixon, A. Doyle, M. Garcia-Hernandez, E. Huala, G. Lander, M. Montoya, N. Miller, L. A. Mueller, S. Mundodi, L. Reiser, J. Tacklind, D. C. Weems, Y. Wu, I. Xu, D. Yoo, J. Yoon and P. Zhang. 2003. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Research, 31:224-228.

Steadman K. J., M. S. Burgoon, R. L. Schuster, B. A. Lewis, S. E. Edwardson and R. L. Obendorf. 2000. Fagopyritols, D-chiro-inositol, and other soluble carbohydrates in buckwheat seed milling fractions. J Agric Food Chem. 48(7):2843-2847.

Wieslander G. 1995. Allergy to buckwheat. Current Advances in Buckwheat Research: 951-955.

Wagner G. P., K. Kin and V. J. Lynch. 2012. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. 131(4):281-5.

The Analysis of Buckwheat Seed and Leaf Transcriptome

Yuhsin Chen, Chieh Wang, Hwang-Bin Chen and Yi-lun Liao
Taitung DARES, COA, Executive Yuan

Abstract

In this study we surveyed the transcriptome of buckwheat seeds and leaves, 143 M and 247 M sequence are read and assembled into 97,200 and 74,544 contigs, the average coverage is 1474 and 3340, respectively. Average length of contigs is 565 and 607 bp, overall sequence assembled is 54 M and 45 M bp. When seed and leaf transcriptome are combined, a total of 106,190 contigs are assembled, in which 69,043 and 56,345 contigs belong to seeds and leaf, respectively. Of all the contigs, 22,857 and 9,889 are exclusively exist in seeds and leaves, respectively. After mixed assembling, 27,258 contigs, 25.6% of all assembled contigs, are considered false assembled as the RPKM value is 0 in both seed and leaf transcriptome. The successfully assembled contigs are then blasted with databases and annotated. The gene ontology (GO) is annotated according to cellular component, biological process and molecular functions. In the project, we are interested in seed properties include seed storage protein, LEA protein, cupin family and allergenic protein. All the above sequence are sorted and retrieved, which would be valuable for future breeding.

Key words: Rutin, Allergenic Protein, Fage1, Fagopyritol.